DOCUMENT RESUME

ED 205 536 TM 810 404

AUTHOR Lam, Tony C. M.
TITLE The Unreliability of Difference Scores for the Measurement of Change: A Paradox?
PUB DATE Apr 81
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Gains; Correlation; Error of Measurement; Hypothesis Testing; Mathematical Models; *Scores; *Statistical Significance; *Test Reliability
IDENTIFIERS *Difference Scores; Interaction (Statistical); *Power (Statistics)

ABSTRACT

The objective of this paper is to examine the relationship between the unreliability of difference scores and the power of tests of significance in an attempt to determine the validity of the paradox for the measurement of change presented by Overall and Woodward: that the power of tests of significance is maximum when the reliability of the difference scores is zero. It is concluded that this paradox is more superficial than real. A negative covariation between power and reliability of difference may occur if subject-by-time interaction variance is significantly greater than measurement error variance, and a positive relationship if the opposite is true. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED205536

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Unreliability of Difference Scores for the Measurement of Change: A Paradox?

Tony C.M. Lam

University of Washington

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. Lam

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 404

Paper presented at the annual meeting of the American Educational Research Association in Los Angeles, April, 1981

Difference scores have been used extensively in educational and psychological research to assess treatment effect or changes either within a single group or across different groups despite the fact that these scores have long been known to be unreliable as a result of additive measurement errors (Bereiter, 1963). Because of this augmented error, one would expect that the rejection of the null hypothesis is less likely, and efforts should be made to increase the reliability of these scores. In contrast to this intuitive deduction, Overall & Woodward (1975) claimed that " in research concerned with differences between experimental treatment groups, the loss in reliability due to calculation of difference scores is not a valid concern because the power of tests of significance is maximum when the reliability of the difference scores is zero "(p.85). The objective of this paper is to examine the relationship between the unreliability of difference scores and the power of tests of significance in an attempt to determine the validity of the paradox for the measurement of change presented by Overall & Woodward (1975, 1976).

The analysis of variance model and the equation for difference score reliability.

In a single group pre-post measures design in which the treatment is introduced between the two testing periods, the model equation is (Winer, 1971, p. 278):

$$X_{ij} = \mu.. + T._j + S_i. + ST_{ij} + E_{ij} \qquad (1)$$

where X_{ij} is the observed score for the ith subject at time j, $\mu..$ is the grand mean, $T._j$ is the effect of the jth time period, $S_i.$ is the ith subject effect $\sim$ independent & $N(0,\sigma_s^2)$, ST_{ij} is the ith subject by jth time effect $\sim$ independent & $N(0,\sigma_{st}^2)$, and E_{ij} is the random measurement error for subject i's score at time j $\sim$ independent & $N(0,\sigma_e^2)$. $S_i.$, ST_{ij}, and E_{ij} are assumed to be independent.

At pretest: $X_{i2} = \mu.. + T._2 + S_i. + ST_{i2} + E_{i2}.$

At posttest: $X_{i1} = \mu.. + T._1 + S_i. + ST_{i1} + E_{i1}.$

$$D = X_{i2} - X_{i1}$$

$$= (T._2 - T._1) + (ST_{i2} - ST_{i1}) + (E_{i2} - E_{i1}).$$

Under the constraint $(ST_{i2} + ST_{i1}) = 0$,

$$D = (T._2 - T._1) + 2(ST) + (E_{i2} - E_{i1}).$$

The expected value of D

$$\xi(D) = \xi(T._2) - \xi(T._1) + 2\xi(ST) + \xi(E_{i2}) - \xi(E_{i1})$$

With the assumptions: $\xi(T._j) = T._j$, $\xi(ST) = \xi(E_{ij}) = 0$,

$$\xi(D) = (T._2 - T._1)$$

$$= \mu_d \quad .$$

The variance of difference scores

$$\sigma_d^2 = \xi(D - \mu_d)^2$$

$$= \xi(2ST + E_{i2} - E_{i1})^2.$$

With the uncorrelated error assumption,

$$\sigma_d^2 = 2(2\sigma_{st}^2 + \sigma_e^2). \qquad (2)$$

Based on the classical test theory's definition of reliability as the ratio of true to observed score variance, the reliability of difference scores is

$$\rho_{dd} = 2\sigma_{st}^2 \; / \; (2\sigma_{st}^2 + \sigma_e^2) \qquad (3)$$

Let π be the reciprocal of σ_d^2

$$\pi = 1 \; / \; [\, 2(2\sigma_{st}^2 + \sigma_e^2) \,] \qquad (4)$$

Under the assumptions of fixed Type I error rate and sample size, and assuming that the difference between population means is nonzero, π is positive monotonic with power and can be used as an index of power (Sutcliffe, 1980).[1]

Relations between reliability of difference scores and power

In examining equations (3) and (4), the following conclusions are apparent,

1) Difference score reliability (ρ_{dd}) is negatively related to measurement error variance (σ_e^2) and positively related to subject-by-time interaction variance (σ_{st}^2).

2) power (π) is negatively related to σ_e^2 and also negatively related to σ_{st}^2.

---

[1] Although the original proposal for this paper presentation was submitted prior to the publication of Sutcliffe's (1980) article, parts of this paper however are influenced by and are indebted to Sutcliffe's (1980) words.

If σ_e^2 is held constant and σ_{st}^2 is free to vary ($0 < \sigma_{st}^2 < \infty$), σ_{st}^2 will covary positively with π; therefore, ρ_{dd} will covary inversely with power and the paradox may be observed (Overall and Woodward, 1976). In the special case considered by Overall and Woodward (1975), when $\sigma_{st}^2 = 0$, $\rho_{dd} = 0$ and $\pi = 1 / 2\sigma_e^2$, "power of tests of significance is maximum when the reliability of the difference scores is zero" (p.85). On the other hand, if σ_e^2 is fixed and σ_{st}^2 is varied ($0 < \sigma_{st}^2 < \infty$), σ_e^2 is negatively related to both ρ_{dd} and π; hence, ρ_{dd} and π will covary positively and the paradox does not exist. In the case where $\sigma_e^2 = 0$, $\rho_{dd} = 1$ and $\pi = 1 / 4\ \sigma_{st}^2$, and the "maximal power for the t test is attained when the reliability of the difference score is also maximal" (Fleiss, 1976).

The two seemingly contradictory conclusions about the form of relationship between reliability of difference scores and power are both correct, and the apparent contradiction is the consequence of imposing different preconditions on σ_{st}^2 and σ_e^2. In other words, the ρ_{dd} and π relation is spurious and the relation depends on the relative contributions of the moderator variables : subject-by-time interaction and error of measurement, or the change in one of these moderator variables relative to the change of the other one. However, the change in σ_{st}^2 may have greater effect on ρ_{dd} than does σ_e^2. This discussion will be expanded upon on page 5. "The 'paradox' has arisen because the indirect associations of power and reliability have been treated as if they were direct" (Sutcliffe, 1980). If the direct relationships between ρ_{dd}, σ_{st}^2, and σ_e^2 are studied, instead of the indirect one between ρ_{dd} and π, the paradoxical power of tests of significance based on differences can be clarified and resolved.

Efficiency of experimental design and its relation to power and reliability

The subject-by-time interaction variance (σ_{st}^2) is the measure of variability of subjects in changing from pretest to posttest time periods. As it was pointed out previously, π increases and ρ_{dd} decreases as σ_{st}^2 decreases. The increase

in power is therefore not a direct function of reducing ρ_{dd}, but rather, it is because subjects in the experiment "happen to be less heterogenous in their differential response to treatment independently of the measuring procedure" (Sutcliffe, 1980). In a randomized blocks repeated measures design or single group pre-post measures design, the reduction of σ^2_{st} is achieved by an increase in subject variance (σ^2_s). Since the subject-by-time sum of squares is equal to the difference between within treatment sum of squares and subject sum of squares ($SS_{st} = SS_w - SS_s$), where SS_w is the experimental sampling error sum of squares in a completely randomized group design. The variation attributable to differences in the subjects means is removed from the experimental error in the randomized blocks repeated measures design (Edwards, 1972, p.260). The experimental design is efficient if large subject variance is attained. Based on this observation, an interesting relation between experimental design and the paradoxical issue can be established. In designing an experiment, considerable effort is usually made in an attempt to increase the efficiency or accuracy of the experiment by reducing residual variability by partitioning from it other additional sources of variation due to individual differences. This reduction of experimental error variance "usually translates into reducing the variability of the true score" (Nicewander and Price, 1978). Therefore, with the increase in efficiency of an experimental design, power is increased but the reliability of the dependent measure is reduced simultaneously.

Correlation between pre and post measures (r_{12}) and subject-by-time interaction (σ^2_{st})

If there is small variability of subjects in changing from the first to the second testing periods, i.e. if σ^2_{st} is minimal, subjects tend to occupy the same relative positions in performance at both pre and posttests. This implies that a high positive correlation between the pre and post measures (r_{12}) will be obtained. This negative relation between σ^2_{st} and r_{12} can be shown by the following observations. First, in the single group pre-post design, $\sigma^2_{st} = \sigma^2_w(1 - r_{12})$, (Edwards, 1972, p.267),

the negative relation is apparent. Second, given the general formula for the reliability of difference scores

$$\rho_{dd} = (\rho_{xx} - r_{12}) / (1 - r_{12}) \qquad (5)$$

where ρ_{xx} is the average of the pre and post measures' reliabilities, it can be seen that ρ_{dd} is negatively related to r_{12}. Since ρ_{dd} is positively related to σ^2_{st}, r_{12} and σ^2_{st} bears a negative relationship.

Associated with equation (5) is the familiar unreliability-invalidity dilemma (Bereiter, 1963), which states that "other things being equal, the higher the correlation between pretest and posttest, the lower the reliability of the difference scores..... the lower the correlation between two tests, the less they can be said to measure the same thing" (p.4). If equations (4) and (5) are compared, it appears that the dilemma is now transformed into the paradox by replacing r_{12} with σ^2_{st} and the concern for validity (r_{12}) with the concern for the power of tests of significance.

In regard to how ρ_{dd} can be improved, Bereiter (1963) pointed out that since increasing the reliability of the pre and post measures will also increase their correlation (r_{12}), changing the values of r_{12} will have greater effect on altering the values of ρ_{dd} than if the values of the pre and post test reliability are changed. Since r_{12} is related to σ^2_{st}, it may be suspected that σ^2_{st} has greater influence than σ^2_e on ρ_{dd}. However, this is only speculative and empirical verification is needed.

The importance of reliability

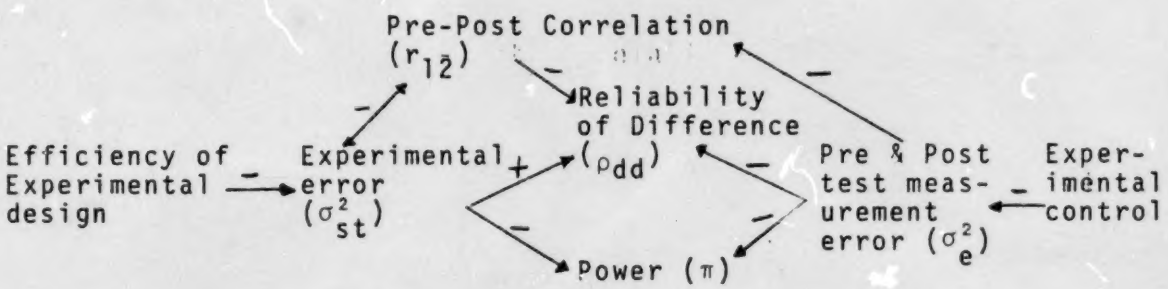
According to Sutcliffe (1980), experimenters should be concerned with the relationship between power and measurement error (σ^2_e), with experimental residual variability (σ^2_{st}) held constant, in relation to reliability of difference. Under this condition, reduction in measurement error at the pre and post measures will increase both power and the reliability of difference. Therefore, Overall & Woodward's (1976) claim that "reduction in reliability associated with calculation of difference scores is not a cause for concern

with regard to testing the significance of treatment effects" (p. 777) is incorrect and misleading. Efforts should be expended to reduce measurement error at both testing periods by tighter experimenter control, e.g. standard experimental procedures, reliable assessment instruments, careful data collection, and so on; or by making more than one observation per subject per treatment as suggested by Sutcliffe (1980); as it appears that sloppy data will not enhance the chance of rejecting the null hypothesis when change scores are used. From the unreliability-invalidity dilemma perspective, Bereiter (1963) concluded that if change scores can be justified as the appropriate meaningful dependent measure, e.g. in measuring attitude change, "the intelligent course would have been to go blithely ahead increasing the reliability of the change scores, letting the test-retest correlation drop to zero if it so happened" (p.5).

Summary

The apparent paradox of the unreliability of difference scores for the measurement of change is more superficial than real. The confusion arises because the two sources (σ^2_{st}, σ^2_e) for power and unreliability of differences were not studied. A negative covariation between power and reliability of difference may occur if σ^2_{st} is significantly greater than σ^2_e, and a positive relationship if the opposite is true.

Other variables were considered in this paper and they can be summarized by the following diagram:



Where the positive and negative signs indicate the direction of the relations, the arrow reflects the sequential relations

of the variables, e.g. the value of the experimental error (σ^2_{st}) is determined by the efficiency of experimental design; and the dual headed arrow between r_{12} and σ^2_{st} implies no precedence of one over the other, i.e. the two variables are related but neither variable affect the other one in the temporal sense.

In a fixed effect one way analysis of variance design, measurement error has been demonstrated to decrease the power of tests of significance (Cleary & Linn, 1969; Sutcliffe, 1980). Based on the argument presented by Sutcliffe (1980), the same conclusion can generally be made when difference scores are used as dependence measure. Contrary to what is implied by the paradox, reliability of measurement is a relevant concern for experimenters who employ a single group pre-post measures design. It is because given a certain amount of residual sampling error in the experiment, higher reliability of difference scores, as a result of low measurement error at both pre and post tests, tends to associate with higher statistical power in testing the difference of means.

References

Bereiter, C. Some Persisting Dilemmas in the Measurement of Change. In C.W. Harris (Ed.) Problems in measuring change. Madison: University of Wisconsin Press, 1963, 3-20.

Cleary, T.A., & Linn, R.L. Error of measurement and the power of a statistical test. The British Journal of Mathematical and Statistical Psychology, 1969, Vol. 22, Part I, 49-55.

Edwards, A.L. Experimental Design in Psychological Research, 1972, Holt, Rinehart & Winston, Inc.

Fleiss, J.C. Comment on Overall & Woodward's Asserted paradox concerning the measurement of change. Psychological Bulletin, 1976, Vol. 83, No. 5, 774-775.

Nicewander, W.A. & Price, J.M. Dependent variable reliability and the power of significance tests. Psychological Bulletin, 1978, Vol. 85, No. 2, 405-409.

Overall, J.E. & Woodward, J.A. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. Psychological Bulletin, 1976, Vol. 83, No. 5, 776-777.

Overall, J.E. & Woodward, J.A. Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 1975, Vol. 82, No. 1, 85-86.

Sutcliffe, J.P. Error of measurement and the sensitivity of a test of significance. Psychometrika, 1958, Vol. 23, No. 2.

Sutcliffe, J.P. On the relationship of reliability to statistical power. Psychological Bulletin, 1980, Vol. 88, No. 2, 509-515.

Winer, B.J. Statistical Principles in Experimental Design, 1971, McGraw-Hill, Inc.